Hourly energy demand and prices: an analysis on risk measures and correlation

Camilo Oberndorfer Mejía Gregorio Pérez Bernal Luisa Toro Villegas

Miguel Valencia Ochoa

August 11, 2022

Abstract

In this paper, diverse methodologies are described and applied to real life data. The main focus of these methodologies is to understand behavior over time, identify outliers, measure dependency within one variable and multiple variables, and attempt to predict future energy variables, like price and demand.

The previously mentioned methodologies will be applied to the hourly energy demand generation in Spain from 2015 to 2019. The hope to this paper is to find a way to predict energy prices using correlation techniques and understand how demand affects supply and vicersa.

Contents

1 Data description

1.1 Original dataset

This dataset included 35048 data of 20 different variables about electric energy generation, demand, price and forecasts of some of these variables.

It included how much energy was generated by every generation method used in Valencia, Spain, how much demand the energy had and the price of the energy at a given time. Every hour of every day from January 1st of 2015 to December 31 of 2018 has its own data consisting on the previously mentioned variables excluding those hours in which the energy generation, demand or price was not measured.

For more detailed reading about the original dataset, it can be found at: [Jha19]

1.2 Modifications

For each method used in this paper, a subset of the original data was chosen according to each specific needs, mainly due to the need to visualize properly in graphs. These subsets are:

- A subset of the data including only the information of the hours of the last month on the dataset which was called data720 including exactly 720 data registries.
- A subset including the data from the first hour of everyday was included in the construction of the control charts, this data includes 1460 data registries and is called data1460.
- A data subset including the data from the first hour of each month was taken, this data set included 48 data and therefore is called data48.

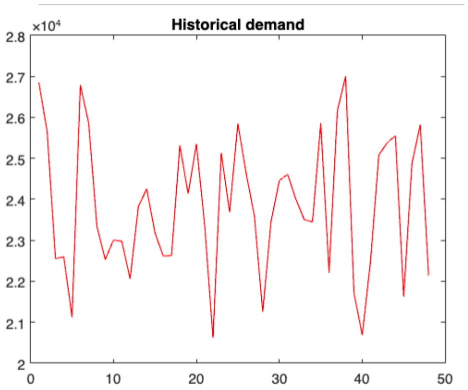
A second dataset that included weather in different regions of Spain during that time. The only data considered was that of Valencia. A problem that was encountered was that several entries of the dataset had categorical values, which means that analysing them was more difficult. This is why methods of Categorical Encoding were used. Basically, the method consists on assigning a numerical value for every non-numerical input. For instance, if the entries regarding climate were: ”Sunny”, ”Rainy”, and ”Cloudy”, each of them would have a numerical value assigned to represent them. It is better to assign them hierarchically, in order for them to have a more realistic way of being analysed.[Yad19]

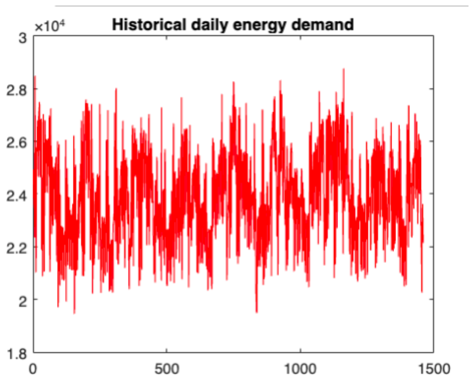
2 Estimation and results

2.1 Historical data

For the historical data, the three main variables of interest (energy demand, energy generation and energy price) were plotted to understand their structure and historical behavior. On the left are the plot for the monthly data and on the right for the daily data of each variable mentioned.

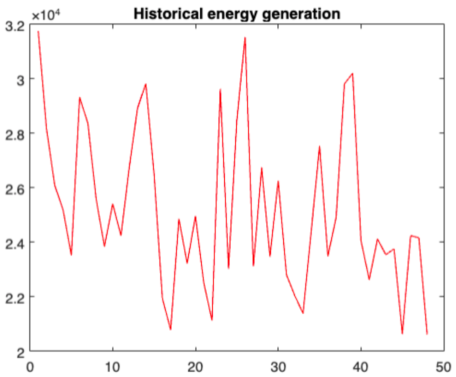
The hourly data's plot is not included because 35000 entries would look too crowded to understand and it provides no extra information. The prediction's data isn't shown either because their behavior is very similar to what they are predicting (specially for demand).

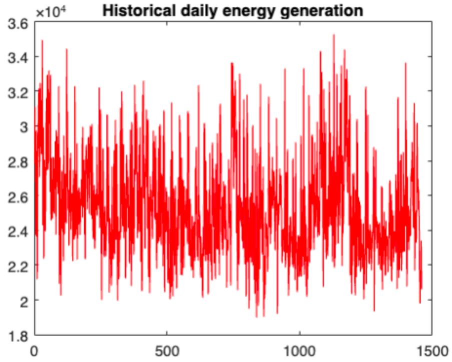




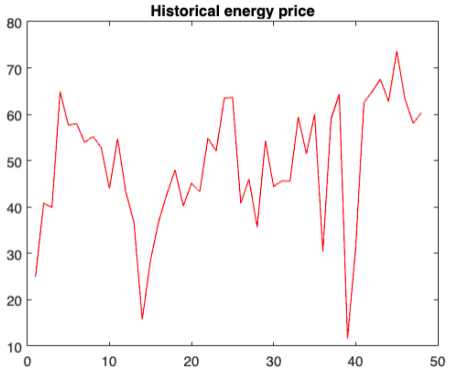
The historical data graph on the left depicts the energy demand of the first hour of every month, as seen, due to the little amount of data, any behaviour or structure can’t be differentiated.

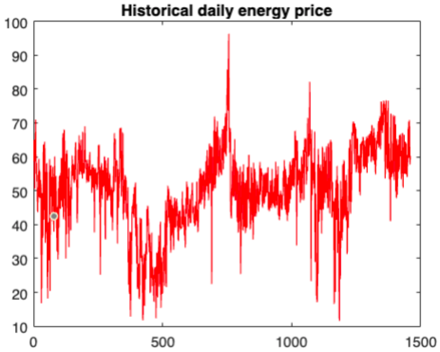
However the graph on the right shows the demand included in the dataset data1460 and shows a more clear behaviour were in over all maintains an oscillatory behaviour, this is probably due to weather seasons, making winter and summer the seasons with more electric energy usage, anyways there is still a lot of noise.





There is shown a clear drop in the generation of electrical energy on both graphs but in the graph to the right it has a more clear oscillatory behaviour similar to the one seen in the demand, later its was realized that the most correlated variable to the generation of energy was the forecast of the demand so its predictable for them to have a similar behaviour.



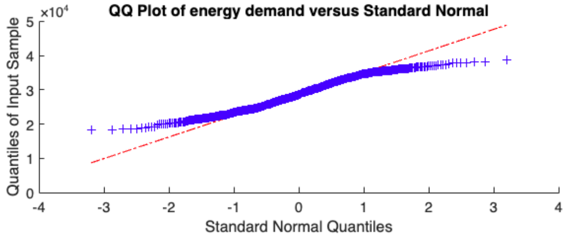


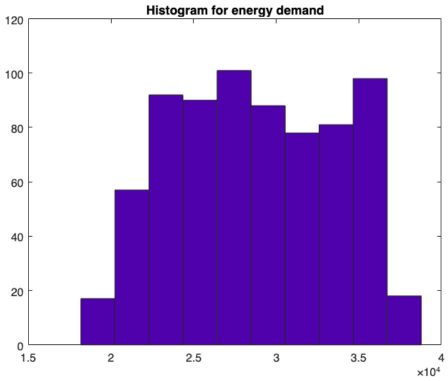
Both graphs show a general slowly rising behaviour on the price of energy with three periods of time were the price changes abruptly, the first drop on the price of energy was probably caused by a previous excess energy generation, this drop itself caused a big demand increase and this later caused the price to go high up again, also a response behaviour is seen as right after the price goes up or down it takes the opposite direction.

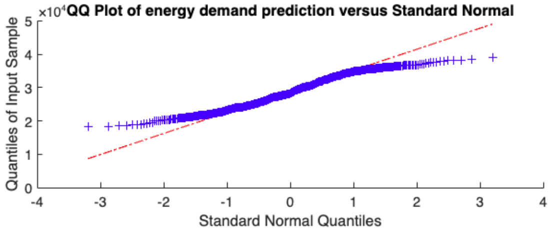
2.2 Identifying underlying distribution

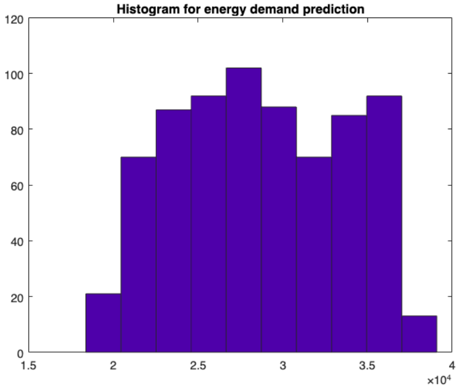
To construct a thorough analysis on any given set of data, a crucial first step is to attempt to understand that way the data is distributed. If the data follow a normal distribution, the following analysis could have a smaller margin of error, but it is rare to see that any real life data follows a normal distribution.

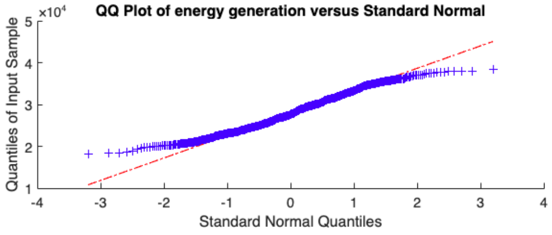
The data’s energy generation histogram and quantile-quantile plot (qqplot) are graphic tools that help identify normality, which are shown in the following figures:

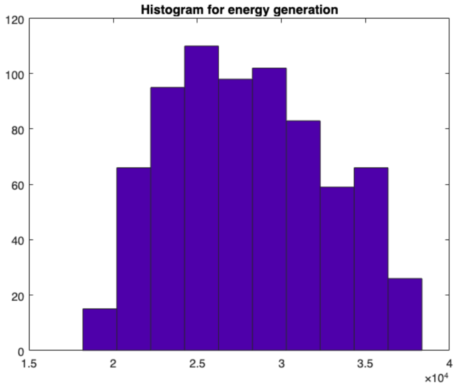












The previous figures are a characteristic example of under-dispersed data, for example a uniform distribution or a beta distribution with parameters close to (1,1) and the qqplot, which compares the theoretic quantiles and sample quantiles, shows an s-shape.

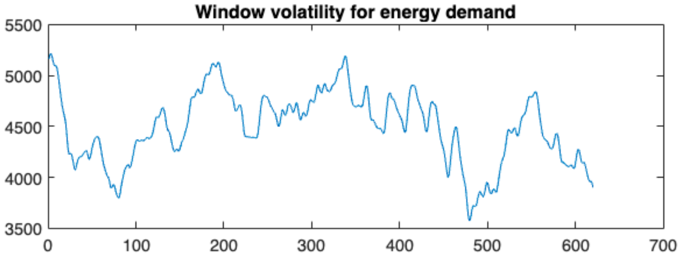
2.3 Stability tests

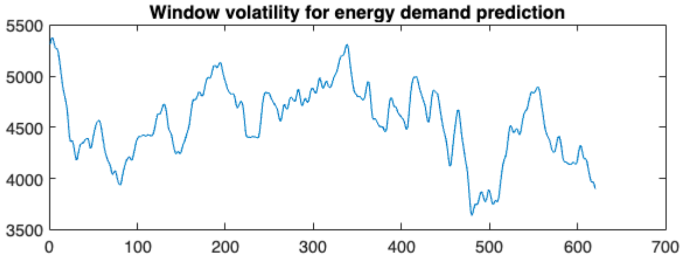
2.3.1 Rolling windows

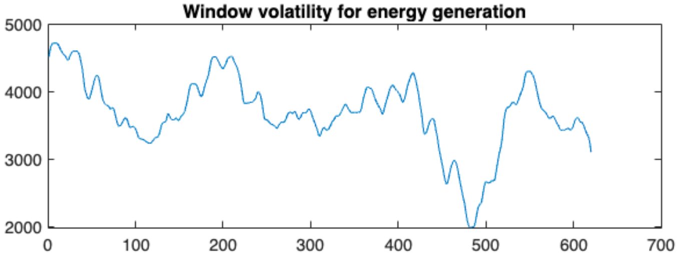
Rolling windows for volatility is a method that allows to capture the behavior of the variables of interest over time, specially considering that in real-life data parameters such as standard deviation don't stay constant but instead, vary over time. Rolling windows are estimated using the following formula [MB06] :

$$\sigma^2 = \sum_{i=t-m+1}^{t} \frac{ri^2}{m}, t = m, m+1, ..., n$$

The following graphs show the standard deviation for each of the variables of interest with windows of $m = 100$.



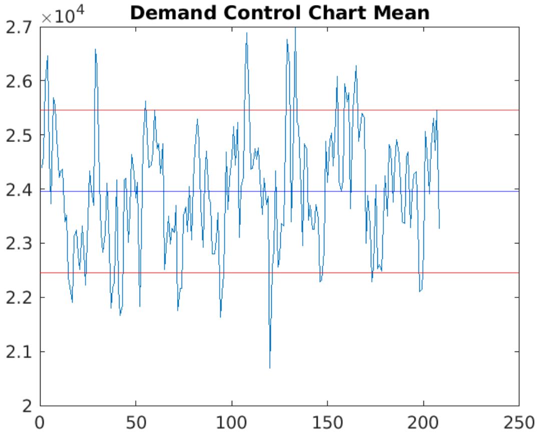


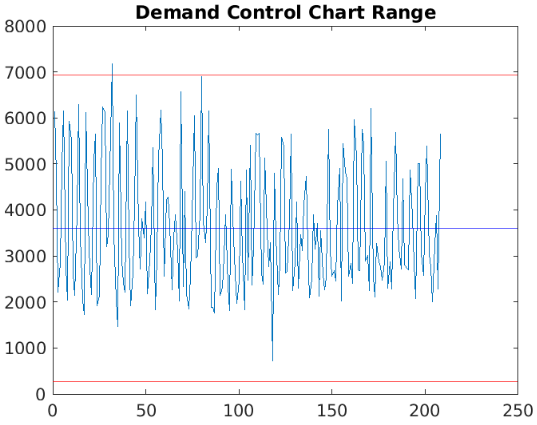


All three variables taken into account show a very similar behaviour of volatility over time and all of them have a usually high volatility this means that even if they are different and a have a different dispersion, it possible for the data of each variable to be dispersed similarly.

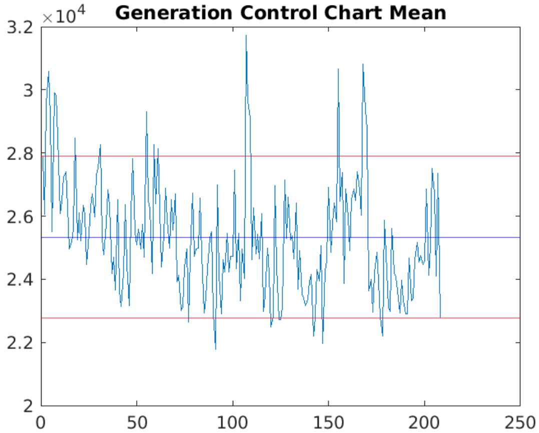
2.3.2 Control Charts

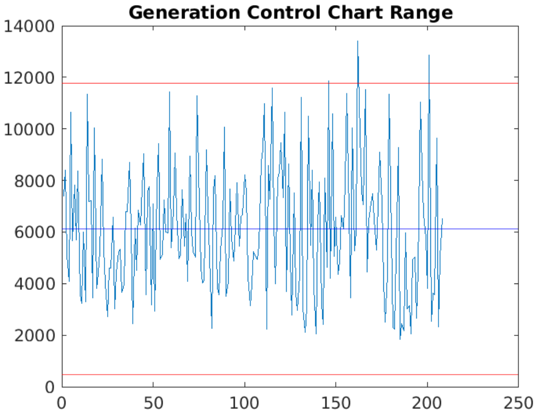
A complete stability analysis is made on an $\bar{x} - R$ chart of the electric energy demand, generation and price over time, samples of 7 hours a week are taken in consideration for the chart, the data1460 is used for this process in each of the three variables taken in count. [GD09]



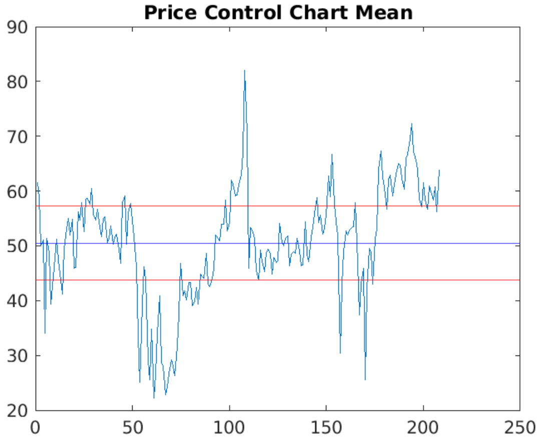


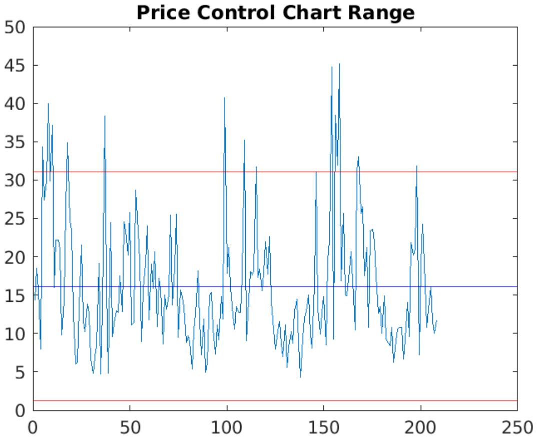
As shown in the previous chart, electric energy demand in Valencia is not under statistic control as there are some outliers with exponential behaviour that surpass the established control limits. The results were predictable as demand has a high correlation with some variables in the dataset but also some other things like special days affect a lot in the quantity of energy demanded.





As shown in the mean control chart, energy generation is a less under control process that has shown a decreasing behaviour due to both, a decrease in resources to produce energy and the increase in the import of electric energy from other cities.





Lastly the price has shown to be the less controlled variable taken in count, this is a predictable outcome due to the low correlation even with the supply and demand so a lot of other unknown variables affect the price and is harder to generate an accurate prediction for the price of energy than to generate a good prediction for the energy demand at a given time.

2.3.3 Exponential smoothing

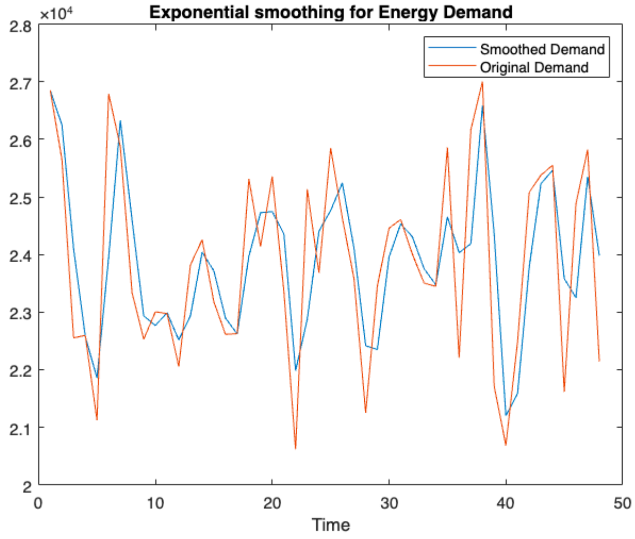
Exponential smoothing is another way of estimating a data series' volatility when it isn't constant. The method consists on generating a smooth function out of raw data, one that describes the general tendency of data over time. The simplest form of an exponential smoothing algorithm is given by the following formula [NB00]:

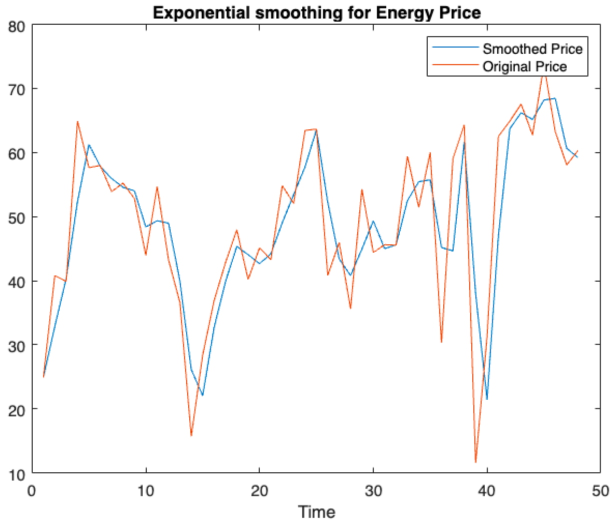
$$s_0 = x_0$$

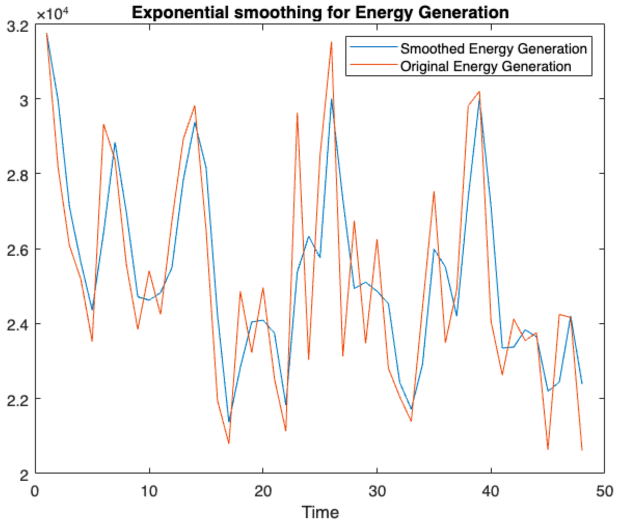
$$s_t = \alpha x_t(1-\alpha)s_{t-1}, t > 0$$

where s_t is the estimation of what the next value of x will be. α is the smoothing factor.

Note that the expression for s_t is a linear combination between the xth observation and and s_{t-1}, meaning that α must be between zero and one. Note that the method works by assigning a higher value for the most recent values. This relation is exponential, thus explaining the name of the method. [NB00] This is exactly why this method is so efficient to make predictions, a process that will be discussed later on the article.



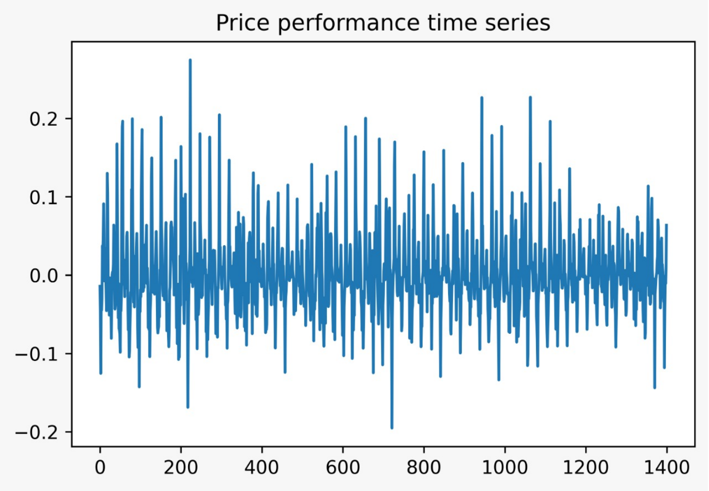




Looking at the following images, where the trends are shown in blue, which show the behavior exponentially smoothed, showing that the most stable variable is energy price. However outliers are present, especially in weeks 12 and 49, where large price drops were present.

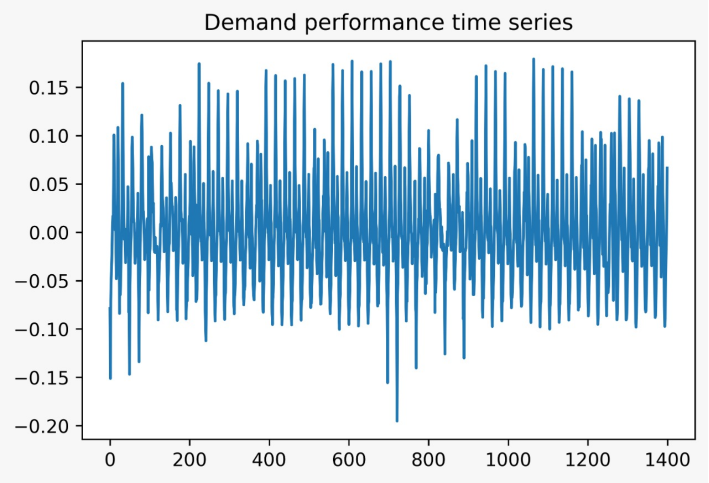
2.4 Price and Demand Performance

Performance is the measurement of whether a value of a variable increased or decreased, regarding their previous observation. Thus calculating the energy and demand performance, can give us an idea of how these variables developed over time, the mean and standard deviation, or risk, of the performance sum up the average performance and its volatility. If the average performance were to be positive it would mean that the variable had an overall increase. If the risk was high, then the variable could change drastically every timestep.



- *Price performance mean:* 0.00267
- *Price performance volatity:* 0.07568

Given that the price performance was positive, it means that there was an overall increase in price, but since the value is relatively close to 0, the change in price isn't really acknowledgeable, the volatility on the other hand even though not very big, it is bigger than the average performance, meaning instability in the performance.



- Demand performance mean: 0.00131
- Demand performance volatility: 0.05156

Almost the same behaviour can be regarded in the demand, the demand slightly increase and was lightly unstable.

2.5 Extreme values and outliers

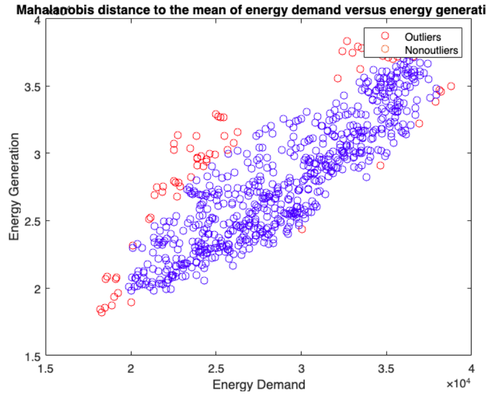
2.5.1 Mahalanobis distance

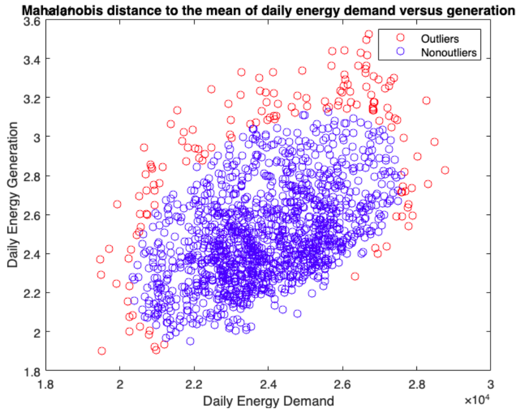
When analyzing data, it is important to detect the presence of outliers, in order to avoid making decisions that would compromise more accurate results.

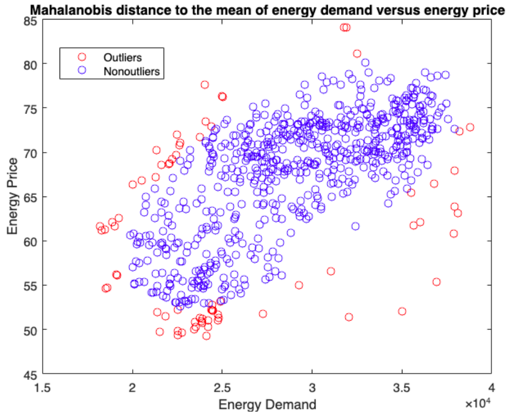
Mahalanobis distance is used to detect outliers in multivariant testing.[Gho19] The distance of an observation $\vec{x} = (x_1, x_2, x_3...x_n)^T$, from a set of observations with means $\vec{\mu} = (\mu_1, \mu_2, \mu_3, ..., \mu_n)$ and a given covariance matrix Σ is defined as:

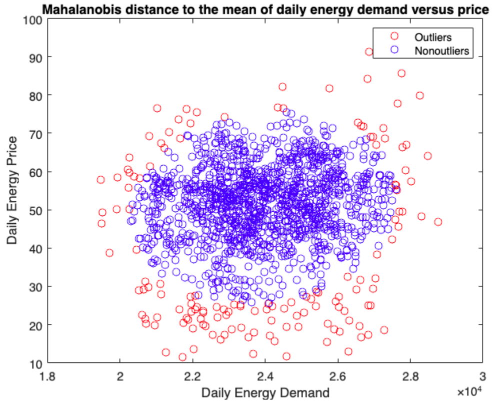
$$D_m(\vec{x}) = \sqrt{((\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu}))}$$

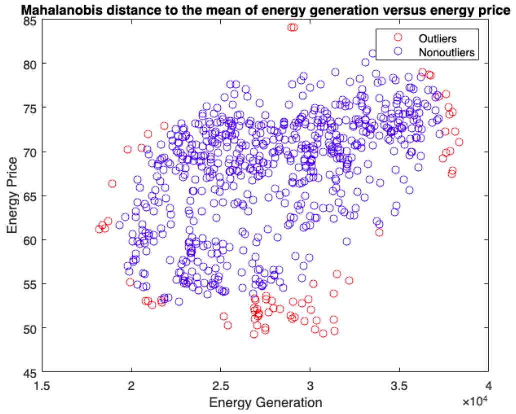
Intuitively, this distance means that from the center, an ellipse is formed, and the entries that are beyond the limits are considered as outliers. This process is important because it helps to detect the stability of a process. The following figures show the Mahalanobis distance in different data.

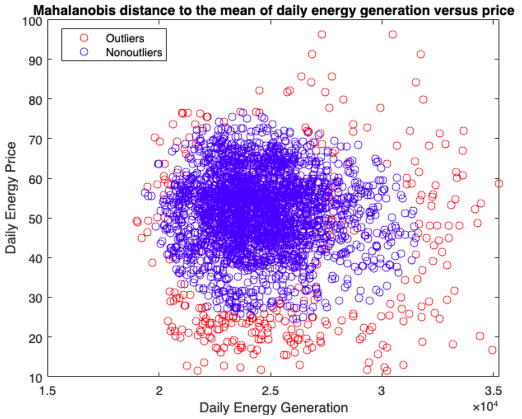










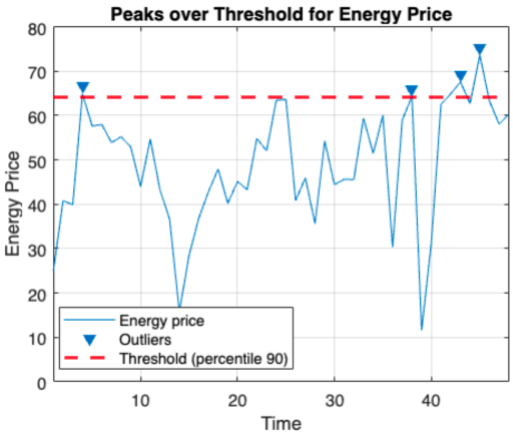


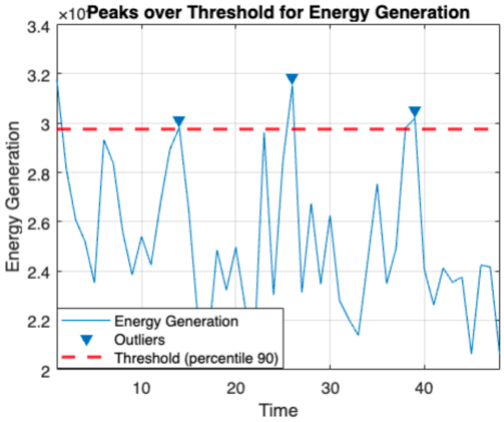
The process that shows the most outliers would be the relation between daily energy generation and its price and also daily energy demand and its price. A reasonable explanation would be the instability of the prices during the entire period. The daily measures for the Mahalanobis distance show a very round shape (almost like a circle) which means that the covariance matrix is similar to the identity matrix, and the Mahalanobis distance is similar to the euclidean distance. This behavior is seen specially in the daily energy generation versus price.

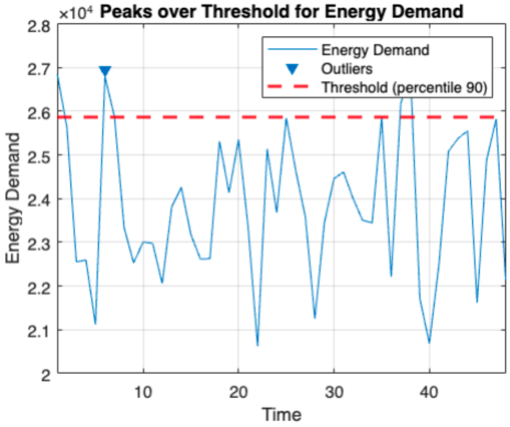
2.5.2 Peaks over Threshold

Peaks over Threshold (POT) identifies the extreme values of the series as those who go over a certain threshold t and these outliers are known as excess returns. In this case, the threshold t es the 90 percentile, which was chosen by the rule of thumb. The graphs for each of the variables of interest POT's are shown bellow. [MB06]

They show that extreme data are distributed amongst all the time series and are in similar positions for each of the plots.







The outliers in all three graphs show now distinct patters and, therefore, the identification of said outliers would only be useful for analysis of "clean" data and not so much in identifying motives for these behaviors.

2.6 Measuring dependency

2.6.1 Autocorrelation Function

Autocorrelation measures the relationship between an observation at time t and its previous observations. It helps detects non randomness in a dataset. Autocorrelation function (ACF) is an specific method for measuring a time series autocorrelation, and measures the correlation between y_t and y_{t+k}, with $k = 0, 1, ..., K$. The difference of time between the previous and current times is referred to as lag k, and its autocorrelation is

$$r_k = \frac{c_k}{c_0}$$

where

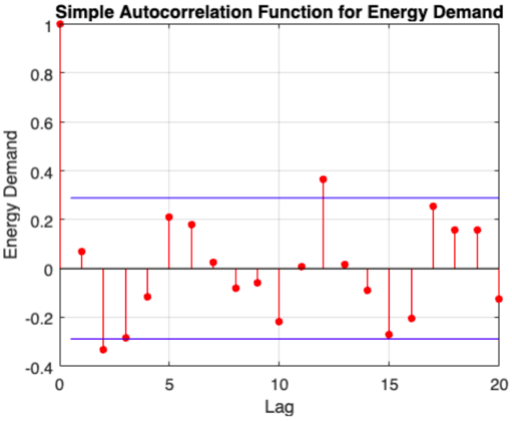
$$ACF(k) = \frac{1}{T} \sum_{t=1}^{T-k} (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})$$

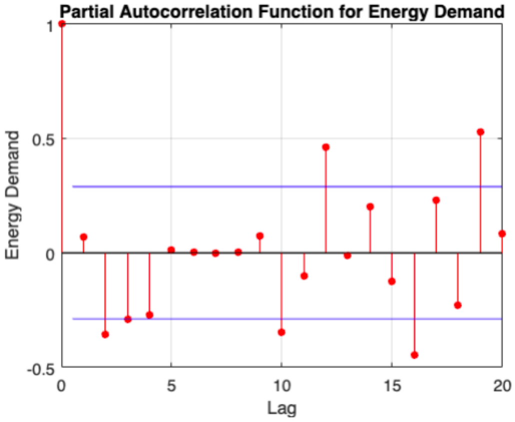
and c_0 is the sample variance.[BJ90]

Partial autocorrelation function PACF is the correlation at lag k is the correlation of values at k time steps apart and all the values in between, different from ACF that only calculates the two values at the intervals' limits.[Dat19]

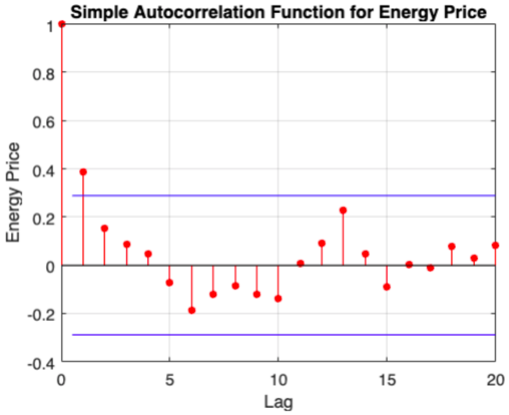
$$PACF(k) = \frac{cov(Y_i, Y_{i-k}|Y_{i-1}, ..., Y_{i-k+1})}{\sqrt{var(Y_i|Y_{i-1}, ..., Y_{i-k+1}) * var(Y_{i-k}|Y_{i-1}, ..., Y_{i-k+1})}}$$

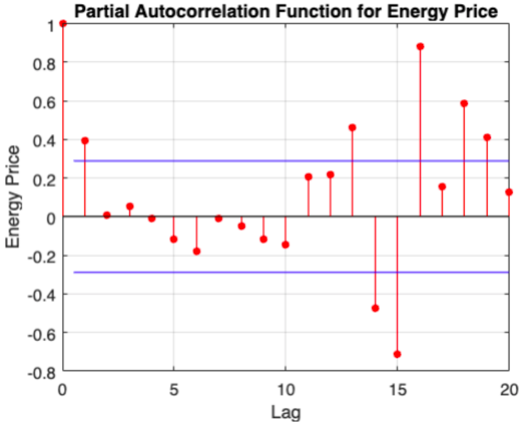
The ACF and the PACF were estimated for data720, and the results are shown bellow. Note that the first bar is always ignored because it measured correlation with itself. A tip for analysing these figures is to determine if the peaks are significant by checking if they cross the thresholds or the confidence interval.



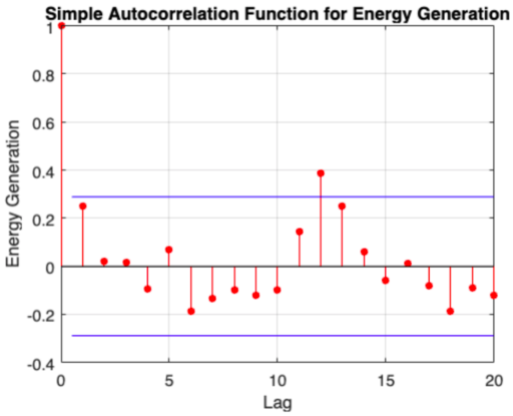


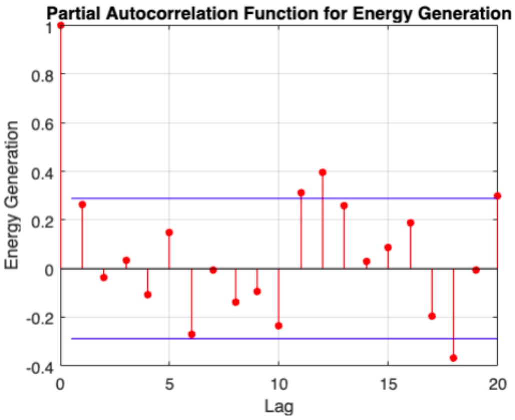
In these first pair of graphs is portrayed the ACF and PACF for energy demand, and they both are very similar, which shows a consistent correlation in the time series. It shows a very small correlation, between $-0.5 \leq R^2 \leq 0.5$ and the peaks that cross the interval cross it by very little, which shows no significant dependency.





In the second pair of graphs, the ACF and PACF are estimated for energy price, and they both are very similar, which means that in the time series there is high correlation, but not in close intervals. The PACF specially shows many peaks way over the threshold and very close to ±1 which means high explicability of itself. This can be explained by the regular price and demand structure were high prices cause a excess in supply and a deficit on demand which then causes the price to go down and vice-versa, this process usually takes some time to take place so is appropriate for the price to behave this way.





The last pair of graphs shows the autocorrelation for energy generation, and both the ACF and PACF are very similar, which means consistency of autocorrelation in the time series. The peaks mostly stay within the interval which signifies low correlation.

2.6.2 Copula

A copula helps detect dependency structures in multivariate data. ”A copula is defined as a function that joins multivariate distribution functions to their one-dimensional marginal distribution functions. It is a multivariate distribution function defined on the unit n-cube $[0, 1]^n$ ”. [Kpa07]

An approach to estimating copulas in a time series with an unknown distribution is to calculate it with the empirical function, which is an empirical measure of the sample and converges with probability 1 to the underlying distribution.

$$\bar{F}_n(t) = \frac{1}{n}\sum_{t=1}^{T-k} IX_i \leq (t)$$

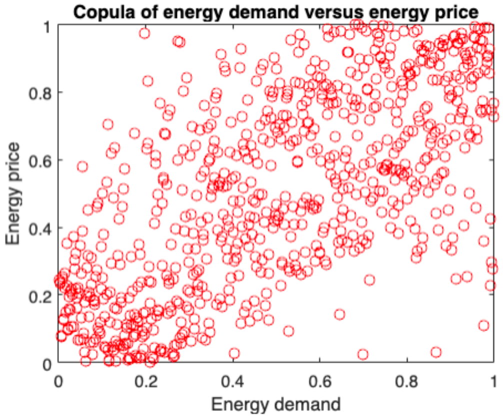
where I is the indicative function.

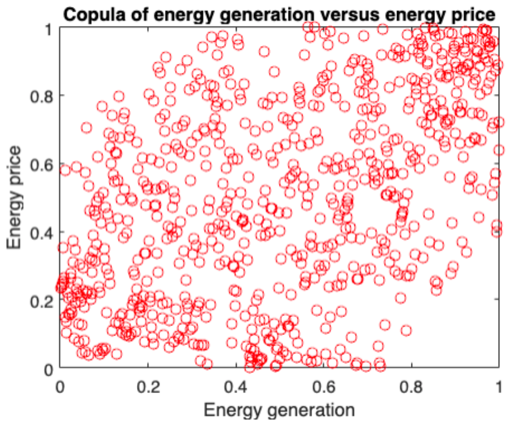
Then, the bivariate empirical copula function is:

$$C_n(u, v) = \frac{1}{n}\sum_{t=1}^{T-k} I[X_i \leq (t)F_n(u), Y_i \leq (t)G_n(v)]$$

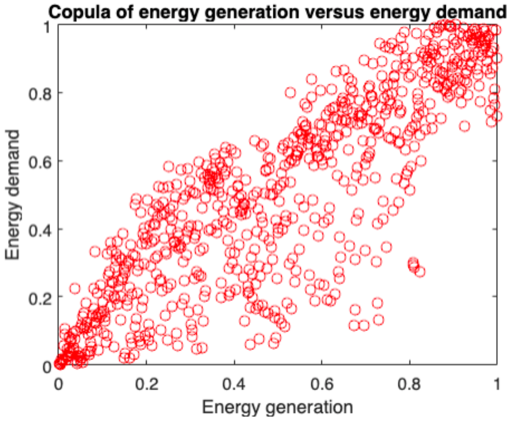
[Kpa07]

The figures bellow show the copulas found for different pairs of the variables of interest:

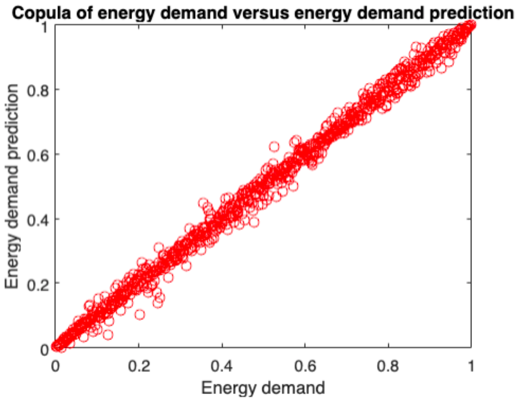


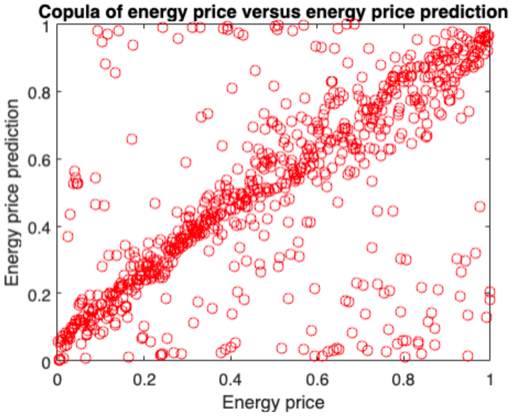


The previous copulas depict a medium structure dependency between the energy demand and the energy generation compared to the energy price in a given time, this means that for making a almost perfect prediction for the price of energy a lot of more different variables are needed.

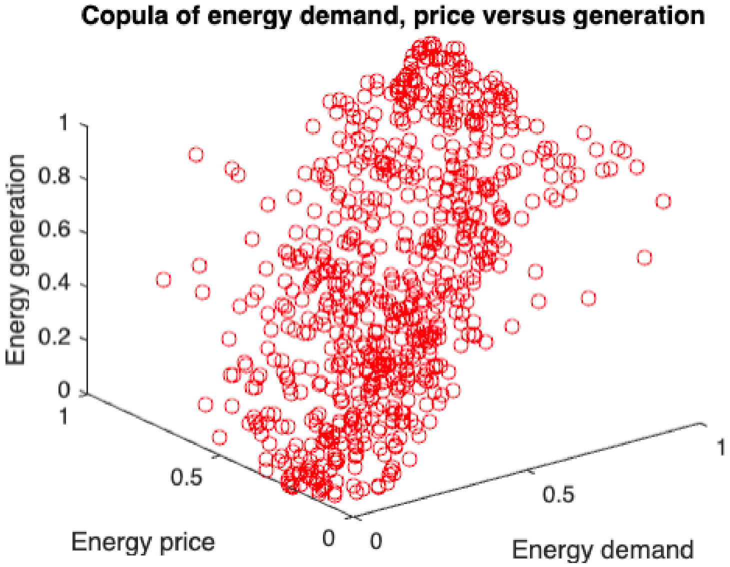


The copula shows us a more structured relation between the energy generation and energy demand very likely making demand one of the most important variables in creating a prediction for the power generation.





Both this copulas show how good are the predictions made in the data sets compared with the actual values of demand and price respectively, as expected, the predictions on demand had less outliers and a better correlation than the predictions done on the price.



In the three dimensional copula of generation, demand and price is seen a better shaped structure which means there is still dependency.

2.6.3 Correlation Coefficient

Correlation is an analysis of two or more different variables which measures their level of relationship. The correlation coefficient R^2 determines how much a variable can be explained in terms of the others, and is measured between $-1 \leq R^2 \leq 1$. The closer the value is to zero, the more independent the variables are from each other and the closer the value is to one, the more dependent the variables are. The sign of R^2 determines the nature of the relationship, positive or negative. [Hay21]

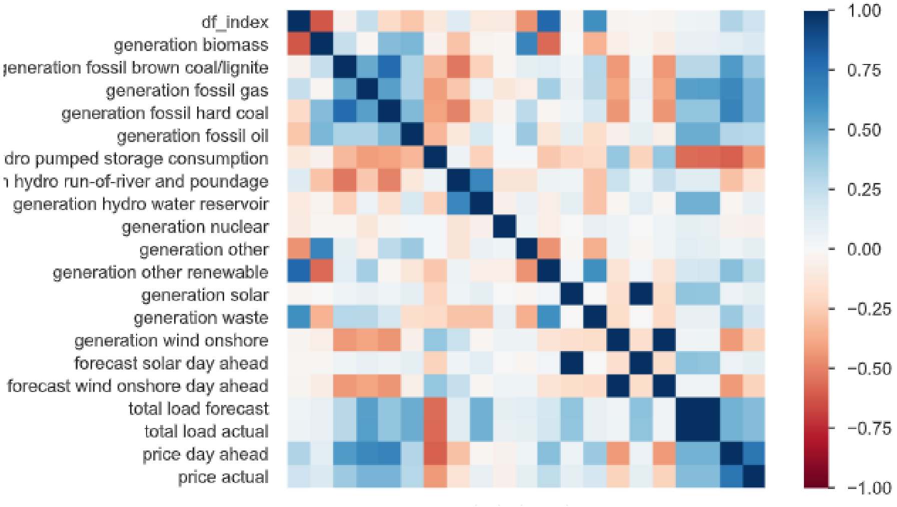
For all three of the graphs bellow, a color scale is used to represent the correlation coefficient. The darker the colors get (red and blue) the more dependency between each of the two variables there is, and the lighter the color is, the less dependency.

Each colored square represents the correlation coefficient between two given variables, and the names on the variables are shown on the left (the are symmetrical with the y-axis).

Pearson's r: The Pearson's correlation coefficient (r) is a measure of linear correlation between two variables. It's value lies between -1 and +1, -1 indicating total negative linear correlation, 0 indicating no linear correlation and 1 indicating total positive linear correlation.

To calculate r for two variables X and Y, one divides the covariance of X and Y by the product of their standard deviations.

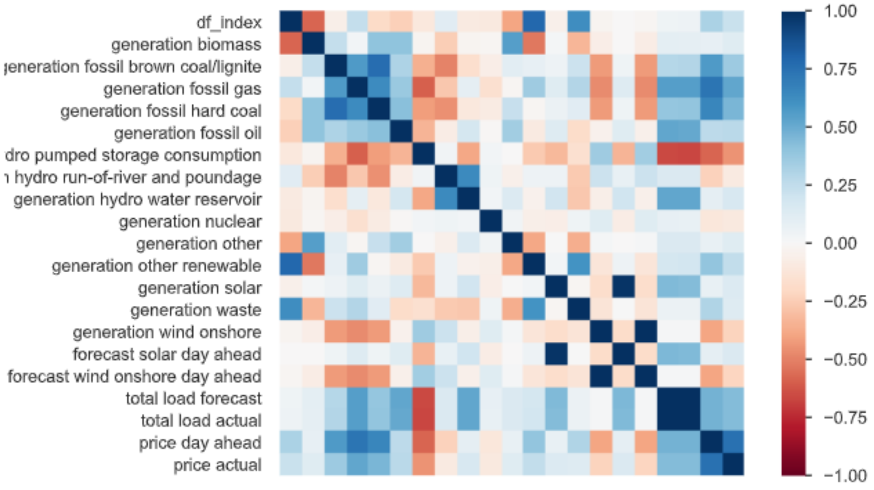
Pearson Correlation Matrix



Spearman's ρ: The Spearman's rank correlation coefficient (ρ) is a measure of monotonic correlation between two variables, and is therefore better in catching nonlinear monotonic correlations than Pearson's r.

To calculate (ρ) for two variables X and Y, one divides the covariance of the rank variables of X and Y by the product of their standard deviations.

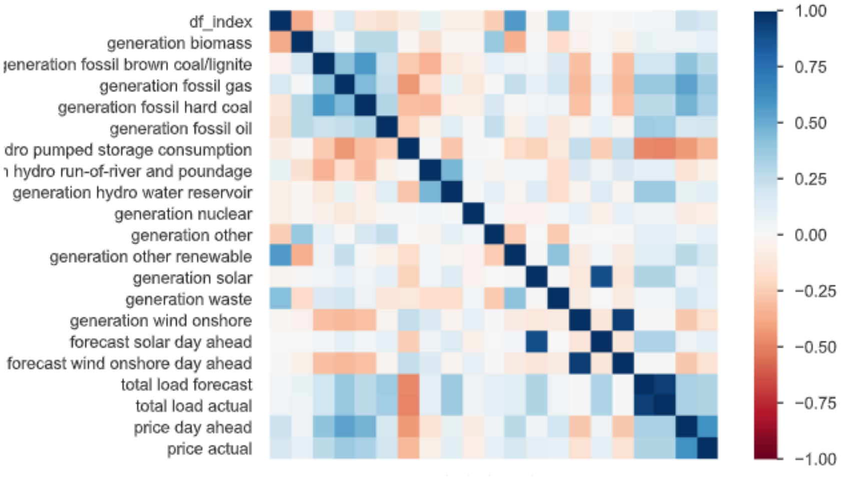
Spearman Correlation Matrix



Kendall's τ: Similarly to Spearman's rank correlation coefficient, the Kendall rank correlation coefficient (τ) measures ordinal association between two variables.

To calculate τ for two variables X and Y, one determines the number of concordant and discordant pairs of observations. τ is given by the number of concordant pairs minus the discordant pairs divided by the total number of pairs.

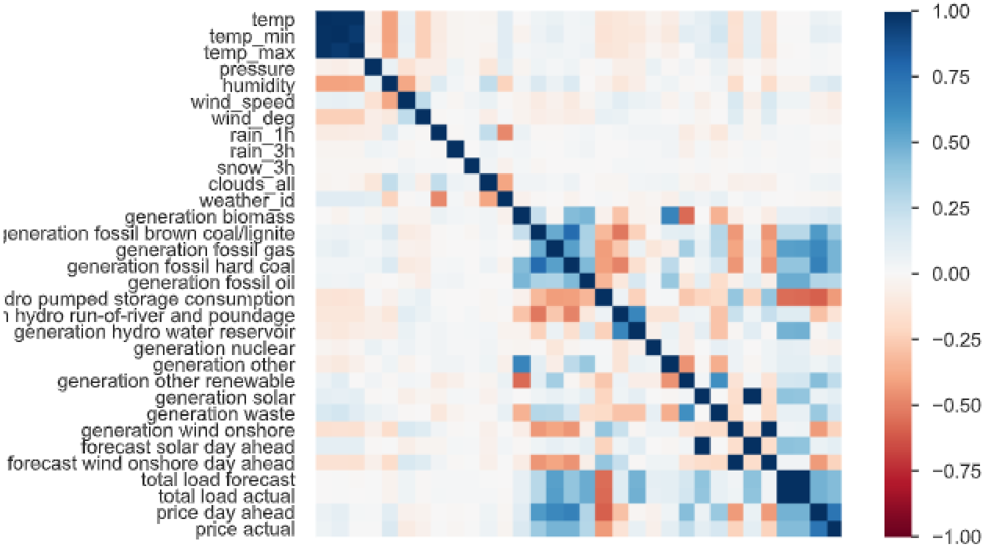
Kendall Correlation Matrix



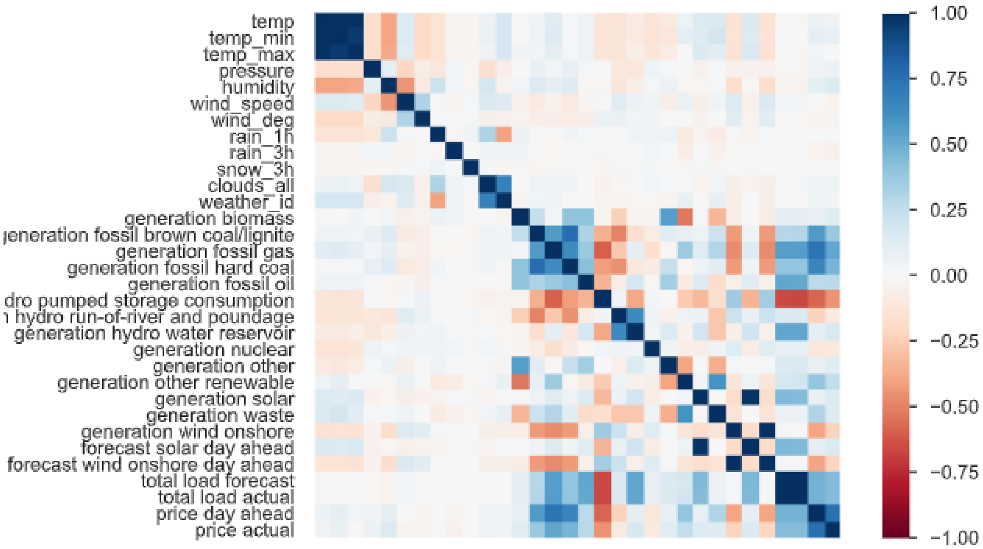
3 Energy correlation with weather

Analysing if the weather had an effect on the production of energy, could give further insight into how to analyse how does the demand and price of energy changes. For this we checked the correlation of every variable with every other variable, between the data set of energy production and weather features in Valencia. Both of this data sets were composed of 35,018 unique timestamps.
It was also important to check the different types of correlation, given that the data may not have a linear correlation, thus the correlation wouldn't be shown with the Pearson formula. Or if the points don't follow an ordinal relation, then both, the Pearson and Kendall formula, wont detect correlation. And the Spearman isn't very accurate with the correlation measures, thus, having the 3 correlation matrix can give an accurate idea of the correlation value and type.

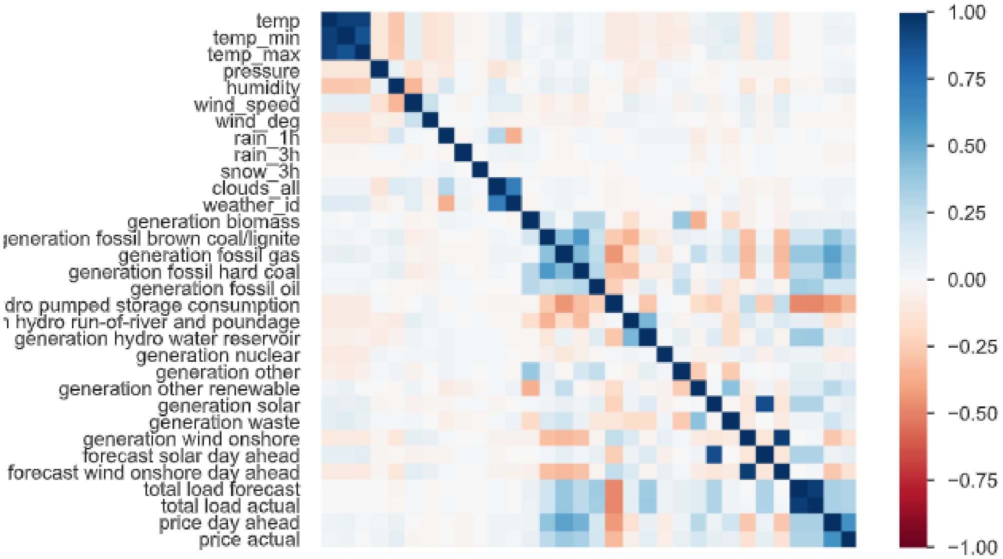
Pearson Correlation Matrix



Spearman Correlation Matrix



Kendall Correlation Matrix



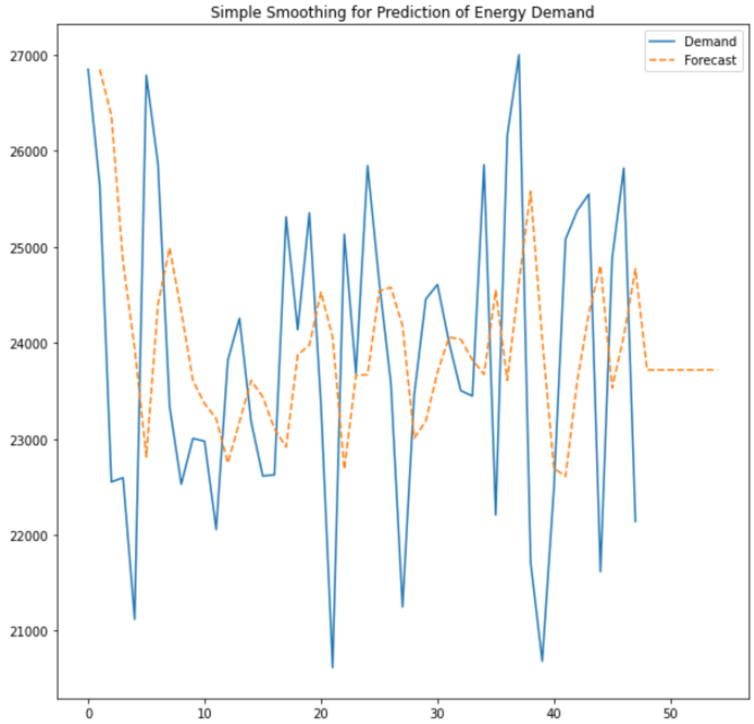
In the three correlation Matrix above we can distinctly observe a lack of correlation between the weather variables and the energy production ones. This can be seen in the fact that mos of the spaces which would correlate a weather variable and an energy production one, are close to being white.
In the energy data set we can clearly observe correlation between many of the variables, these likely have a linear relation, these can be seen in the fact that both the Pearson and Spearman coefficients have a higher correlation values than the Kendall one.

4 Prediction

In this section, three different models for prediction are implemented. Electrical energy prediction is fundamental for decision making in big companies to balance demand and supply, and storing surplus energy is costly and difficult.

4.1 Exponential Smoothing prediction

The first model is a very basic prediction, which is based on the recent past. This model is not very accurate for long-term prediction but its helpful for short-term. Its prediction uses a level, which is an average value around which the time series varies, and it's also useful to ”clean” outliers from a time series, as mentioned in section 2.3.3 [Van19a]



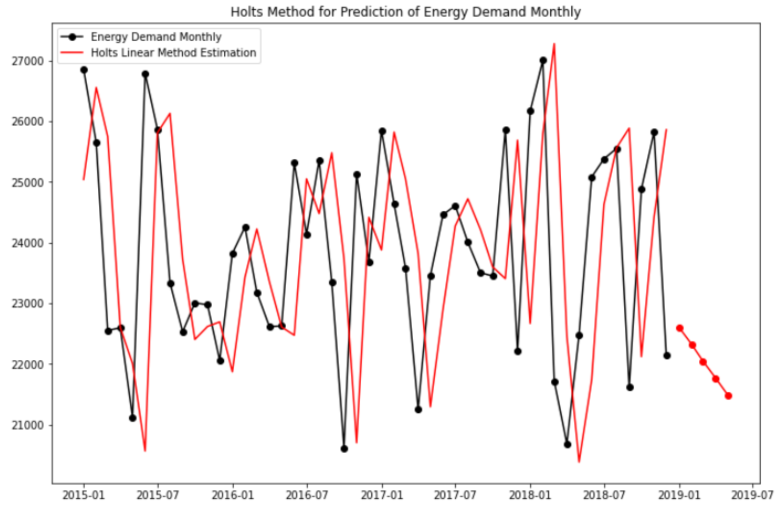
- Mean Absolute Error: 1643.115
- Root Mean Squared Absolute Error 1978.34

As explained above, exponential smoothing predicts how a certain asset will behave as it assigns importance to variables depending on the time in which they occurred, giving exponentially more importance to recent events in order to predict more accurately.

The graphs show that price will stay relatively stable for the following five weeks. Now, the algorithm implemented was the one that was described below. [Van19b] Consider that this method does not give the most accurate predictions, however, it can be very useful in order to predict trends.

4.2 Holt's Model Prediction

Holt's two-parameter model, also known as linear exponential smoothing, is a popular smoothing model for forecasting data with trend. Holt's model has three separate equations that work together to generate a final forecast. The first is a basic smoothing equation that directly adjusts the last smoothed value for last period's trend. The trend itself is updated over time through the second equation, where the trend is expressed as the difference between the last two smoothed values. Finally, the third equation is used to generate the final forecast. This is a different approach towards exponential smoothing in order to predict seasonal change. [Swa00]



- Mean Absolute Error: 1821.5444
- Root Mean Squared Absolute Error 2327.79

This figure shows an implementation of Holt's method used to predict monthly energy demand. Note that the prediction written in red states shows that energy demand is likely to decrease on the future. However, note that this method is not prefect, due to the fact that just at the end of the time measured, the difference between the real energy demand and Holt's estimate is approximately 4000 units, which is not a very accurate prediction.

4.3 Auto Regressive Method for Time Series Forecasting

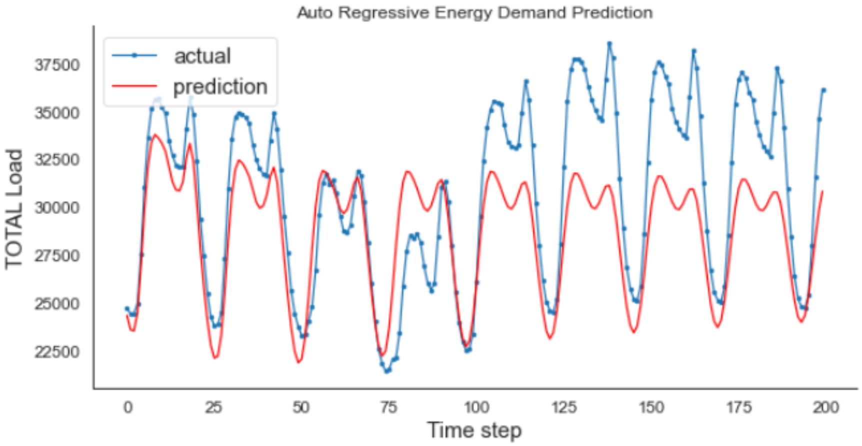
In statistics and time series analysis, an autoregressive model is a prediction method in which a value from a time series is regressed on previous values from the same time series. [SR19] Given y measured on a time y_t will be explained with the previous measurement y_{t-1} as shown below:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t$$

in which β_0 refers to a constant β_1 is the parameter of the model and ϵ_t refers to white noise.

These models can have any number of states, as long as it is a whole number smaller than the number of parameters of the model. [Unind] A generalized way of the method for an order p is as follows:

$$y_t = c + \epsilon_t \sum_{i=1}^{p} \beta_i y_{t-1}$$



- Mean Absolute Error: 0.14494
- Root Mean Squared Absolute Error 0.17461

The following method was applied in energy demand, in order to predict how it would change. Note that the time-step was changed in order to get a more smooth result. The results gotten show that demand is actually going to increase in the timestep given, which could be pondered with Holt's result in order to get an accurate prediction, however, it is recommended to do this when the timestep is the same, in order to get more accurate results.

5 Implementation

Code at [lto21]

6 Future work

- Implement a more accurate model for prediction.
- Dive deeper into copulas and correlation analysis.
- Recognize structure in the outliers identification.

References

[BJ90] George Edward Pelham Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Inc., USA, 1990.

[Dat19] Sachin Date. Understanding partial auto-correlation. *Towards Data Science*, 2019.

[GD09] Humberto Gutiérrez Pulido and Román De La Vara Salazar. *Control Estadístico de la Calidad y Seis Sigma*, volume Segunda Edición. McGraw Hill, 2009.

[Gho19] Hamid Ghorbani. Mahalanobis distance and its application for detecting multivariate outliers. *Facta Universitatis Series Mathematics and Informatics*, 34:583, 2019.

[Hay21] Adam Hayes. Correlation definition. *Investopedia*, 2021.

[Jha19] Nicolas Jhana. Hourly energy demand generation and weather. *Kaggle*, 2019.

[Kpa07] Tchilabalo Kpanzou. Copulas in statistics. *University of Stellenbosch*, 2007.

[lto21] ltorov. *Hourly-energy-demand-and-prices-an-analysis-on-risk-measuresand-correlation*. Github, 2021.

[MB06] Luis Fernando Melo Velandia and Oscar Becerra Camargo. Medidas de riesgo, características y técnicas de medición: una aplicación del var y el es a la tasa interbancaria de colombia. *Banco de la República, Gerencia Técnica*, 2006.

[NB00] Mohamed N. Nounou and Bhavik R. Bakshi. Chapter 5 - multiscale methods for denoising and compression. In Beata Walczak, editor, *Wavelets in Chemistry*, volume 22 of *Data Handling in Science and Technology*, pages 119–150. Elsevier, 2000.

[SR19] Varan Singh Rohila. Hourly energy demand time series forecast. *Kaggle*, 2019.

[Swa00] P. M. Swamidass, editor. *Forecasting models, Holt'sHOLT'S FORECASTING MODEL*, pages 274–274. Springer US, Boston, MA, 2000.

[Unind] Penn State University. *Autoregressive Models*, chapter 14. Eberly College of Science, n.d.

[Van19a] Nicolas Vandeput. Simple exponential smoothing for time series forecasting. *Towards Data Science*, 2019.

[Van19b] Nicolas Vandeput. Simple exponential smoothing in python from scratch. *Towards Data Science*, 2019.

[Yad19] Dinesh Yadav. Categorical encoding using label-encoding and one-hot-encoder. *Towards Data Science*, 2019.